Kannpos-Kannada Parts of Speech Tagger Using Conditional Random Fields

K.P. Pallavi and Anitha S. Pillai

Abstract Parts Of Speech (POS) tagging is one of the basic text processing tasks of Natural Language Processing (NLP). It is a great challenge to develop POS tagger for Indian Languages, especially Kannada due to its rich morphological and highly agglutinative nature. A Kannada POS tagger has been developed using Conditional Random Fields (CRFs), a supervised machine learning technique and it is discussed in this paper. The results presented are based on experiments conducted on a large corpus consisting of 80,000 words, where 64,000 is used for training and 16,000 is used for testing. These words are collected from Kannada Wikipedia and annotated with POS tags. The tagset from Technology Development for Indian Languages (TDIL) containing 36 tags are used to assign the POS. The n-gram CRF model gave a maximum accuracy of 92.94 %. This work is the extension of "Parts of Speech (POS) Tagger for Kannada Using Conditional Random Fields (CRFs).

1 Introduction

Parts of Speech (POS) tagging is a process in which each word of a sentence is tagged with appropriate syntactic label such as noun, verb, adjective, preposition, and so on [1]. The syntactic label that represents these lexical categories is known as a tag. For example, NN label represents a common noun tag, and VMF label represents a finite verb tag.

Example Kan: rAma (N_NNP) shAlege (N_NN) hOda (V_VMF).
En: Ram (N_NNP) went (V_VMF) to (PSP) school (N_NN).

A tool that tags appropriate POS of each word in a given document is known as POS tagger. A POS tagger can be developed using various techniques like rule based techniques, transformation based techniques, Machine Learning

K.P. Pallavi (✉) · A.S. Pillai
Department of Computer Applications, Hindustan University, Chennai 603103, India

N.R. Shetty et al. (eds.), *Emerging Research in Computing, Information, Communication and Applications*, DOI 10.1007/978-81-322-2553-9_43

(ML) techniques, and example based techniques and so on. Rule based techniques have been written using syntactic rules, whereas Machine Learning techniques use probabilistic models and stochastic grammar. In this paper, a supervised machine learning technique called Conditional Random Fields (CRFs) is discussed in developing POS tagger for Kannada.

Kannada is one of the south Indian Languages belonging to the Dravidian family. It is the native language of the Karnataka state and spoken mostly in the southern parts of India. Even though Kannada has 50.8 million speakers[1] all over the world and is the official language of the Karnataka[2] state, research in computational linguistics in Kannada is still lagging. One of the main reasons for research to lag in computational linguistics for Kannada is due to its rich morphological and agglutinative nature [2]. It is an agglutinating language with suffixes and nominative-accusative syntax. It also follows Subject-Object-Verb (SOV) constituent order [3]. A word in Kannada could be comprised of 8 suffixes, where the word could either be a single word or combination of words consisting of root words and suffixes. It could also be made up of two or more words, which leads to ambiguity. Pre-processing rules to split those kinds of words have been written according to orthography of a language. Most of the other ambiguities are solved by using a fine grained TDIL hierarchical tagset. Tagset is one of the key elements of text processing techniques like Chunking, Named Entity Recognition, POS. POS tagset have been proposed for few Indian Languages by Technology Development for Indian Languages (TDIL). For Kannada POS, several tagset exist like TDIL [4], IIIT Hyderabad,[3] Bhuvaneshwari C. Melinamath [5], Vijayalakshmi patil's [6], Shambhavi [2] and Antony [7]. Along with the tagset, corpus, a collection of large raw data, plays an important role in training and testing POS tagger. For effective Natural Language Processing (NLP) results, corpus should be from the same domain where NLP application will be prefer. For example, the biomedical POS trained corpus gives better accuracy only for the biomedical datasets. Similarly, the generic area/domain trained corpus gives good results for general datasets. The POS tagger used a large corpus of 80,000 tokens collected from Kannada Wikipedia was trained using Conditional Random Fields (CRFs).

Conditional Random Field is a Supervised Machine Learning (SML) technique. SML techniques are widely being used for text processing, which considers labelled datasets for training. CRFs offer a unique combination of properties by discriminating trained models for sequence segmentation and labelling [8]. Labelled dataset used for training this POS tagger contains manually annotated POS tags with pre-processed corpus words and some extracted linguistic features of those words. This model achieved a competitive accuracy.

Second section highlights the related work done in this field. In the third and fourth sections, challenges and proposed parts of speech tagger for Kannada are

[1] http://timesofindia.indiatimes.com/india/Indiaspeak-English-is-our-2nd-language/articleshow/5680962.cms?referral=PM.

[2] http://en.wikipedia.org/wiki/Kannada.

[3] http://ltrc.iiit.ac.in/nlptools2010/files/documents/POS-Tag-List.pdf.

discussed respectively. Finally, in the last two sections, the paper presents results achieved and concludes with the future work.

2 Related Work

Parts of Speech tagger was developed enormously in English and other European languages. Earlier POS taggers were mostly based on rule based and supervised machine learning techniques. One among those English POS taggers by Toutanova et al. [9] set the benchmark by achieving an accuracy of 97.25 %. After this, the researchers turned their head towards unsupervised machine learning techniques. An unsupervised grammar induction task was demonstrated experimentally that the universal POS categories generalize well across language boundaries and gave competitive accuracy without relying on gold POS tags [10]. Apart from these kinds of POS taggers, Dipanjan et al. [11] approached with token and type constraints for cross-lingual part-of-speech tagging for 8 languages, which reduced error rate by 25 % when compared to the prior state of the art [12]. These taggers were not working well for social media data like twitter data. Kevin et al. [13] has developed a POS tagger for twitter data which gave 90 % accuracy and they believe that their approach can be applied to address other linguistic analysis needs as they continue to rise in the era of social media and it's rapidly changing linguistic conventions.

Recently, A HMM based POS tagger for Hindi and Marathi was developed by Nisheeth Joshi et al. [14] and Jyothi Sing [15]. Hindi POS tagger achieved 92 % accuracy, using Indian Language (IL) POS tagset. Jayabal Ganesh et al. [16] achieved a precision of 87.74 % for Tamil POS tagger using only a small set of POS labelled suffix context patterns. Biswa Ranjan Das et al. [17] has got an accuracy of 81 % for Odia POS tagger using artificial neural networks. A rule based Graphical User Interface tool has been developed by using Netbeans IDE 7.1 for Manipuri by Kh Raju Singha et al. [18]. The 35 rules were framed using 1500 lexicons and achieved an accuracy of 92 %. A Malayalam POS tagger using statistical approach with the Hidden Markov Model following the Viterbi algorithm is described by Jisha P Jayan et al. [19].

Antony et al. [7] proposed a POS tagger for Kannada using Support Vector Machine (SVM) using hierarchical tagset consisting of 30 tags with a compatible accuracy of 86 % and Shambhavi et al. [2, 20] randomly tried with Maximum Entropy (Maxent), Hidden Markov Model (HMM) and CRF. CRF proved better than Maxent and HMM with an accuracy of 84.58 % for 51,269 training data and around 2932 of test data tokens collected from Enabling Minority Language Engineering (EMILLE) corpus. Siva Reddy et al. [21] and Mallama et al. [22] developed POS taggers which gave competitive accuracies for Kannada using Hidden Markov Model and Decision Trees.

In the above Kannada POS papers there are many challenges which were not attempted. Some of the challenges are described in next section.

3 Challenges

All Indian languages are agglutinative in nature [3] where many suffixes or other morphemes are added to the base of a word. Kannada is one such south Indian Languages with 15 vowels, 25 consonants. Developing a Parts Of Speech tagger for Kannada is challenging due to its rich morphological and highly agglutinative nature. Rich morphology occurs with word inflections and are called word level ambiguities. For example, "AguMtukanaMte ((ಆಗುಮತುಕನಮತೆ))" is a word where nouns "AguMtuka" joined with demonstrative "aMte". Consider another example "kELugara ((ಕೇಳುಗರ))" is a word where verb "kEL", it becoming noun after joining with "ugara". Similarly, many ambiguities exists at sentence level also. A sentence level ambiguity includes word tagged as adjectives, postpositions, nouns and adverbs, depending on the POS. For example, ugra ((ಉಗ್ರ)). The method to overcome these kind of challenges are proposed in the following section.

4 Proposed Kannpos

Kannada POS tagger was developed using CRFs for the corpus contained various articles. The corpus was annotated manually using TDIL tagset. Tagset, CRFs, corpus used and annotation method are explained below.

4.1 Tagset

TDIL tagset is unique when compared with the other tagsets. TDIL [23] POS schema is based on W3C XML Internalization best practices and one to one mapping table for all the labels used. It consist of 11 main categories Noun, Pronoun, Demonstrative, Verb, Adjective, Adverb, Postposition, Conjunction, Particles, Quantifiers and Residuals of POS which are classified into 29 level 1 and 5 level 2 sub categories. It proposed different levels of verb category, where other tagsets missed this part. For example, the level 1 verb tags are—main, verbal and auxiliary. Level 2 verb tags are—finite, non-finite, infinitive and gerund. We used TDIL tagset for tagging corpus.

4.2 Corpus

The source of data used in this POS tool was collected from Kannada Wikipedia articles from different domains like sports and eminent personalities. Around

80,000 words were collected, annotated manually, trained and tested in 80:20 ratio. The POS tagger is tested on this corpus and, also on newspaper data.

4.3 Pre-processing Rules

The corpus was pre-processed manually. Pre-processing rules were written to separate symbols and punctuation marks from words. In Indian languages, the symbols and punctuations are written with words as in Manipuri [18]. The words are also written together in Kannada, those words were orthographically split using python script.

Examples

1. Kan: mAyeyeMdhare → mAye + eMdhare
   En: illusion means → illusion + means
2. Kan: rajyadalliruva → rajyadalli + iruva
   En: In the state → state

In example 1, the noun "mAye" suffixed to the demonstrative "eMdhare" and in example 2, the non-finite verb "iruva" suffixed to the common noun "rajyadalli".

4.4 Annotation

Annotation is a complex task and consumes enormous time, compared to corpus collection. Corpus was tokenized through python programs before annotation. Tokenization is splitting the sentences into each word called as tokens. Tokens were arranged in column wise and one blank line was given between sentences before annotation. The 6700 words were annotated manually. Another 73,300 words were tagged using that base engine. Tagged words were validated, and any incorrectly tagged words were corrected manually. Annotated corpus was trained using CRFs.

4.5 Learning Techniques

Conditional Random Fields is a supervised machine learning technique. According to Lafferty et al. [8], "It is a framework for building probabilistic models to segment and label sequence data. Conditional random fields offer several advantages over hidden Markov models and stochastic grammars for such tasks, including the ability to relax strong independence assumptions made in those models. Conditional random fields also avoid a fundamental limitation of maximum entropy

Markov models (MEMMs)". One unique feature of CRF is, it enables the incorporation of arbitrary local features in a log-linear model [13].

In this work, CRFs has been used to develop POS tagger. Since, CRFs is a supervised learning method, rules were written to training the model. Rules were designed using unigram and bigram for tagging. Unigram rules were based on the list look up approach, where bigrams generate conditional probability [24] rules based on the given features to overcome the difficulties faced from Unigram. Bigram rules determined the correct tag when the same word is tagged with different tags in training data. The rules were constructed using words $\{w_{-2}, w_{-1}, w_0, w_1, w_2\}$, features $\{f_0, \ldots, f_{12}\}$. Although the rules were framed using words and features, the annotated tags were used to predict the tag of each word in a training data set. The sequence of words, features extracted from the words and annotated tags were used as input to training CRF model, whereas the output was the POS tags tagged for each word in a sequence of sentences. The features used to frame rules are briefed below.

4.6 Features

Features plays a very important role for CRFS. Identified features includes words and extracted linguistic information of words. Here are the features in detail.

Word: Parts Of Speech exists within the context boundary limit of 5 words. So word feature with window size 5 was used. That means, if the current word was w0, then previous two words were w − 1, w − 2 and next two words were w + 1, and w + 2 are considered.

Case markers: In Kannada, there are 8 case markers for nouns. They are listed with examples in Table 1 and those were identified by comparing with word suffixes. The case markers are not limited to nouns. These can be the inflections of pronouns, verbal nouns and verbs as well. In the training set, we identified 978 words coming with case markers, but they were not nouns. Those words are listed below in Table 2 and some examples of those kind of words are given in Table 3.

Last character: Many tokens/words occur with the same set of last characters. Words with same last character were identified.

Prefix: Almost all the pronouns starts with similar kind of letters. So, prefix features helped in identifying the pronouns easily. Apart from pronouns many other words starts with similar kind of characters. Those were identified easily with the combination of other features.

Tense: Tense markers play a very important role in verbs. There are 60 finite verbs occurring with case markers. This ambiguity was solved by identifying tense and

Table 1 The list of case markers of Kannada

Vibhakti (Case)	Pratyagalu (Markers)	Examples
prathama (Nominative) dvithiya (Accusative) thrithiya (Instrumental) chaturthi (Dative) paMchami (Ablative) shasti (Genitive) saptami (Locative) Sambodana (Vocative)	ು (u) ನ್ನು(annu) ಇಂದ (iMda) ಗೆ (ge) ಕೆ (ke) ಇಕೆ (ike) ದೆಸೆಯಿಂದ (deseyiMda) ಅ (a) ಅಲ್ಲಿ (alli) ಏ (E) ಓ (O) ಆ (A)	Huduganu pustakavannu batteyiMda tayige ramanige Danakke mudukana deseiMda akkana americadalli gurugaLE siteyOramA

Table 2 List of tags with the number of tags which are nouns, but coming with case markers

Tags names	Number of tags
Demonstratives	251
Verbal nouns	188
Pronouns	173
Postpositions	137
Particles	88
Finite verbs	60
Conjunctions	29
Quantifiers	19
Adverbs	18
Adjectives	13
Auxiliary verbs	2

png markers for verbs. In Kannada grammar, we have 3 tenses past, present and future. The case marker for past tense is "da", future tense markers are "uva" and "va" and the past tense marker is "utta". Sometimes the present tense marker is used as future tense. So, according to the modern linguistics/grammar, Kannada morphology has only two tenses, past and non-past [3]. Other than these, some tense markers which occurred frequently in our training set are listed in Table 4.

PNG (Person, Number and Gender markers): PNG marker gives the information on nouns from verbs. These are always coming with verbs. Common PNG markers in Kannada are: Lu, nu, ru, ge and gi.

Table 3 Examples for words which are not nouns, but coming with case markers

Words with Vibakthi's but not nouns	Vibakthis	Tags of the words	Words with Vibakthi's but not nouns	Vibakthis	Tags of the words
ಆ	ಆ	DM_DMD	ಬೇರೂರಬೇಕೆ	ಕೆ	DM_DMQ
ಇಂದ	ಇಂದ	DM_DMR	ರಿಗೆ	ಗೆ	DM_DMD
ಗಳಿಗೆ	ಗೆ		ರೊಂದಿಗೆ	ಗೆ	
ಹೇಗೆ	ಗೆ	DM_DMD	ಅಗಾಗ್ಗೆ	ಗೆ	DM_DMD
ಎಂಬುದಕ್ಕೆ	ಕೆ	DM_DMD	ಈಬಗ್ಗೆ	ಗೆ	DM_DMD
ಅನ್ನು	ಅನ್ನು		ನಿಗೆ	ಗೆ	
ಯಾಕೆ	ಕೆ	DM_DMD	ನಮ್ಮೊಂದಿಗೆ	ಗೆ	DM_DMD
		DM_DMD	ಬಾರಿಗೆ	ಗೆ	DM_DMD
		DM_DMQ			PR_PRP
					PSP

Table 4 Tense markers with examples

Tense markers	Examples
"ುತ್ತೇ"	ಉತ್ತರಿಸುತ್ತೇವೆ
"ುತ್ತಿ"	ಬಳಲುತ್ತಿದ್ದ
"ುತ್ತಾ"	ಬರೆಯುತ್ತಾರೆ
"ುವ"	ಬರೆಯುವ
"ಿದ"	ನೋಡಿದ
"ಗೊಂಡು"	ಕೈಗೊಂಡು
"ಕೊಂಡು"	ತೆಗೆದುಕೊಂಡು
"ತಿದ್ದ"	ಎದುರಿಸುತ್ತಿದ್ದ
"ಕ್ಕೀಡು"	ಬಂಧನಕ್ಕೀಡು

Examples Kan: ashwini (N_NNP) shalege (N_NN) hodaLu(V_VM_VF)
En: Ashwini (N_NNP) went (V_VM_VF) to (PSP) school (N_NN)

In the above example, Lu is the feminine marker which identified Ashwini. L gives information for a second person and also person is in one number.

Negation markers: The verbs which are ending with illa and alla belong to the negation category. Verbs with negation markers were treated as negations, along with negation markers. Some nouns include negation information. Those were separated from negation markers.

Examples thinnuvudilla, maduvudilla

Digits: Most commonly, English number symbols are found in Kannada script. But in few cases only Kannada number used. So, to avoid the confusion both English and Kannada number symbols were grouped to form a Digit Feature. The suffixes attached with the number like raMdu, ralli, nE makes the task easy. Suffixes were treated as demonstrators.

Examples Kan: 12raMdu
En: On 12th

Here, the preprocessed rule was written to separate 12 and raMdu. Then 12 tagged as digit and suffix raMdu tagged as demonstrator.

Symbols: Symbol feature helped in identifying the symbols as well as surrounding words.
Punctuations: The comma, full stop, question mark and colon together forms punctuation feature. The sentence breakers occurs after a full stop, colon and question mark. Including comma, all punctuation helps in finding surrounding words.
Foreign words: Non-Kannada script comes into this group.

All the above features were played important role in increasing the precision of POS tagger. It is reported in results section.

5 Results

The POS tagger tool was experimented on sentences taken from various articles. These sentences consisted of 80000 words, out of which first 64000 were taken as training data and the remaining 16000 words were taken as test data. The test data contained 40 % of words that were not part of training data. The POS tagger tool tagged 16000 test data and the results are presented below in measure of precision. The precision is a ratio of the number of tagged words (*True Pasitive*) and the number of positive responses (*True Pasitive* + *False Pasitive*) as shown below.

$$Precision(P) = \frac{TruePositive}{TruePositive + FalsePositive}.$$

Table 5 Fivefold validation results

Folds	Precisions
1st	91.45
2nd	87.56
3rd	90.57
4th	90.8
5th	88.24
Average	89.724

Table 6 Randomized fivefold validation results

Folds	Precisions
1st	92.14
2nd	92.1
3rd	91.94
4th	92.1
5th	91.8
Average	91.998

Table 7 Sample results

Words	Tags
ಕೇಂದ್ರ	N_NN
ಸರಕಾರದ	N_NN
ಯೋಜನೆಗಳನ್ನು	N_NN
ರಾಜ್ಯಗಳ	N_NN
ಮೇಲೆ	PSP
ಹೇರುವ	V_VM_VNF
ಬದಲು	V_VM_VINF
ಪ್ರತಿಯೊಂದು	N_NN
ರಾಜ್ಯದ	N_NN
ಅಗತ್ಯಕ್ಕೆ	N_NN
ಅನುಗುಣವಾಗಿ	N_NN
ಮತ್ತು	CC_CCD
ಅವುಗಳಿಗೆ	PR_PRP
ಸರಿಹೊಂದುವ	V_VM_VNF
ಯೋಜನೆಗಳನ್ನು	N_NN
ರೂಪಿಸುವುದಕ್ಕೆ	V_VM_VNF
ಪ್ರಧಾನಿ	N_NN
ನರೇಂದ್ರ	N_NNP
ಮೋದಿ	N_NNP
ಒಲವು	N_NN
ವ್ಯಕ್ತಪಡಿಸಿದ್ದಾರೆ	V_VM_VF
.	RD_PUNC

where, True Positives means the system tag is same tag as the gold tag.

False Positives means the system tag is different tag compared to the gold tag.

POS tagger achieved a precision of 87.78 % for 16000 tokens. These tokens were divided into 80:20 ratio for fivefold validation. The fivefold experiment gave the precision between 87 and 91 % shown in Table 5.

The sentences were randomized to training and test again. We have achieved the precisions between 91.8 and 92.94 % for fivefold validation and it is shown in Table 6. We experimented on a new data set of 2,000 words, which is collected from the Kannada daily newspaper and gave precision of 94.7 %. The sample results are shown in Table 7.

6 Conclusion

In this paper, an approach for Kannada POS tagger with hierarchical tagset has been proposed. It has been trained on 64,000 words, and tested on 16,000 unseen words. These 80,000 words were collected from Kannada Wikipedia. They were annotated manually with the help of base engine which was developed on 6,700 words. An annotated words were added with the 12 linguistic features and saved in a training file along with annotated POS tags. The features were designed based on case markers, prefixes, suffixes, tense markers, verb suffixes, png markers, numbers, symbols, foreign words and punctuations based on the linguistic of the language. The linguistic rules were constructed based on permutations and combination of those words with 5 window size. With this tagger, we achieved a competitive accuracy of 92.94 % compared to the other existing works.

This work can be extended in developing NLP applications like Chunker and Named Entity Recognizers for Indian languages. It can be extended by applying unsupervised algorithms to identify the hidden features in the corpus.

References

1. Pallavi., Pillai, A.S.: Parts Of Speech (POS) Tagger for Kannada using conditional random fields (CRFs). In: National Conference on Indian Language Computing (NCILC 2014) 1st to 2nd Feb 2014
2. Shambhavi, B.R., Kumar, R.: Kannada Part-Of-Speech Tagging with Probabilistic Classifiers. Int. J. Comput. Appl. **48**(17), 0975–888, June 2012
3. Sridhar, S.N.: Modern Kannada Grammar. Lordson Publishers Pvt. Ltd., Delhi (2007) (First published in 1990 by Routledge, England under the title *Kannada* in the Descriptive Grammars Series edited by Bernard Comrie. Reprinted in 2007 by Manohar Publishers and Distributors, 4753/23 Ansari Road, Daryaganj, New Delhi 110 002. ISBN 81-7304-767-7)
4. Unified Parts of Speech (POS) Standard in Indian Languages—Draft Standard–Version 1.0, Department of Information Technology Ministry of Communications and Information Technology, Govt. of India

5. Melinamath, B.C.: Improvement over IL-POST tagset for Kannada. Int. J. Comput. Sci. Eng. (IJCSE) **3**(3), 179–186 May 2014
6. Patil, V.F.: Designing POS Tagset for Kannada, Linguistic Data Consortium for Indian Languages (LDC-IL), Organized by Central Institute of Indian Languages, Department of Higher Education Ministry of Human Resource Development, Government of India, March 2010
7. Antony, P.J., Soman, K.P.: Kernel based part of speech tagger for Kannada. In: International Conference on Machine Learning and Cybernetics (ICMLC), vol. 4, pp. 2139–2144, IEEE (2010)
8. Lafferty, J., McCallum, A., Pereira F.C.N.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the International Conference on Machine Learning (ICML-2001)
9. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1, pp. 173–180, Association for Computational Linguistics (2003)
10. Spitkovsky, V.I., Alshawi, H., Chang, A.X., Jurafsky, D.: Unsupervised dependency parsing without gold part-of-speech tags. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1281–1290, Association for Computational Linguistics (2011)
11. Täckström, O., Das, D., Petrov, S., McDonald, R., Nivre, J.: Token and type constraints for cross-lingual part-of-speech tagging. Trans. Assoc. Comput. Linguist. **1**, 1–12 (2013)
12. Li, S., Graça, J.V., Taskar, B.: Wiki-ly supervised part-of-speech tagging. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1389–1398, Association for Computational Linguistics (2012)
13. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for twitter: annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, vol. 2, pp. 42–47. Association for Computational Linguistics (2011)
14. Joshi, N., Darbari, H., Mathur, I.: HMM based POS tagger for Hindi. In: Proceeding of 2013 International Conference on Artificial Intelligence, Soft Computing (AISC-2013) (2013)
15. Bagul, P., Mishra, Archana, Mahajan, Prachi, Kulkarni, Medinee, Dhopavkar, Gauri: Rule Based POS Tagger for Marathi Text. Proc. Int. J. Comput. Sci. Inf. Technol. (IJCSIT) **5**(2), 1322–1326 (2014)
16. Ganesh, J., Ranjani Parthasarathi, T. V. Geetha, and J. Balaji. "Pattern Based Bootstrapping Technique for Tamil POS Tagging." In *Mining Intelligence and Knowledge Exploration*, pp. 256–267. Springer International Publishing, 2014
17. Das, B.R., Patnaik. S.: A novel approach for odia part of speech tagging using artificial neural network. In: Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013, pp. 147–154. Springer International Publishing (2014)
18. Singha, Kh.R., Singha, Ksh.K.B., Purkayastha, B.S.: Developing a part of speech tagger for Manipuri. Int. J. Comput. Linguist. Nat. Lang. Process. **2**(9) Sept 2013
19. Jayan, Jisha P., Rajeev, R.R.: Parts of speech tagger and chunker for malayalam: statistical approach. Comput. Eng. Intell. Syst. **2**(2), 68–78 (2011)
20. Shambhavi, B.R., Ramakanth, K.P., Revanth, G.: A maximum entropy approach to Kannada part of speech tagging. Int. J. Comput. Appl. **41**(13), 9–12 (2012)
21. Reddy, S., Serge S.: Cross language POS taggers (and other tools) for Indian languages: an experiment with Kannada using Telugu resources. Cross Ling. Inf. Access **11** (2011)
22. Reddy, M.V., Hanumanthappa, M.: POS Tagger for Kannada Sentence Translation. Int. J. Emerg. Trends Technol. Comput. Sci. (IJETTCS) **1** (2012)

23. Department of Information Technology Ministry of Communications & Information Technology Govt. of India. Unified Parts of Speech (POS) Standard in Indian Languages—Draft Standard–Version 1.0
24. Che, W., Wang, M., Manning, C.D., Liu, T.: Named entity recognition with bilingual constraints. In: HLT-NAACL, pp. 52–62 (2013)